Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 6 | June 2017

RECENT DEVELOPMENTS IN AGRICULTURE USING DATA MINING TECHNIQUES

V. Manimekalai [1] | R. Suresh [1]

[1] Assistant Professor, Department of Computer Technology, Dr. N. G. P Arts & Science College, Coimbatore 641048.

ABSTRACT

This survey covers some very recent applications of data mining techniques in the field of agriculture. This is an emerging research field that is experiencing a constant development. In this paper, we first present two applications in this field in details; in particular, we consider the problem of discovering problematic wine augmentations at the early stages of the process, and the problem of predicting yield production by using sensor data information. Secondly, we briefly describe other problems in the field for which we found very recent contributions in the Scientific literature.

KEYWORDS: Clustering, Fermentations, Cross-validation, k-means algorithm, Bi-cluster.

1. INTRODUCTION

Two years ago, one of the authors of this survey co-authored a book named "Data Mining in Agriculture" [15]. The book gives a wide overview of recent data mining techniques, and it also presents several applications in the field of agriculture, as well as in other related fields, such as biology. We will give particular attention to two recent works in which the authors of this paper are directly involved, and then we will briefly mention some other applications that looked to us to be the most interesting to report. The survey is organized as follows. In Section 2, we will present an analysis performed on datasets of wine fermentations with the aim of predicting problematic fermentations at the early stages of the process. In Section 3, we will consider the problem of predicting yield production, in which state-of-the-art GPS technologies are employed in connection with site-specific and sensor-based treatments of crops.

2. STUDYING WINE FERMENTATIONS

Wine is widely produced all over the world. There exist different types of wine, which depend by different factors, and especially by the origin of the grapes that are employed in the production. A common point for all wines is the fermentation process, in which the sugar contained in the grapes is transformed in alcohol. This is a very delicate process. When producing wine industrially, indeed, large quantities of wine may get spoiled because of a problematic fermentation process, causing losses to the industry. In order to overcome to this issue, a prediction of the problematic wine fermentations could be attempted, so that an enologist can interfere with the process in time for guaranteeing a good fermentation. In order to monitor wine fermentation processes, metabolites such as, for example, glucose, fructose, organic acids, glycerol and ethanol can be measured. However, analyses are usually limited to data that are obtained within the first 3 days of fermentation. Naturally, this is done in order to learn about a possible problematic fermentation at the beginning of the process. Fermentations can be divided in 3 classes: the first class contains normal fermentations, while the second and the third one contain the problematic ones. In particular, the second class contains fermentations which are slow, in the sense that they can bring the wine to the end of the production. Finally, the third class contains stuck fermentations,

Given a certain time t during the fermentation processes, measurements taken at time t can be grouped together in order to form clusters. A clustering technique might indeed define clusters that are related to normal or problematic fermentations by exploiting the inherent characteristics of the data. Naturally, due in large part to the time-variable nature of the fermentation process, fermentations can be assigned to different clusters for a different t. In these studies, the k-means algorithm [12] was employed for finding clusters of data points, where the number of clusters k was arbitrarily set to 5.

Samples are organized on the columns of a matrix A, and therefore measurements of the same compound taken from different fermentations, but at the same time t, can be found on the rows of A. For each compound and each time t, there is a specific feature in A. A bi-cluster is a sub matrix defined by a subset of samples and a subset of features contained in A. As a consequence, a bi-clustering of A is a partition of A in disjoint bi-clusters, whose rows and columns cover the ones in A, and therefore it gives a relation between samples and features in A [3]. In order to find a consistent bi-clustering, a fractional optimization problem with binary variables can be defined, whose aim is to select the features that are actually relevant for the representation of the sample. This optimization is NP-hard [14]. A heuristic algorithm [13] can be used for the solution of this problem. This bi-clustering technique was able to find some interesting information regarding the compounds that are monitored during the fermentation process [15]. Figure 1 shows the basic Data mining process.

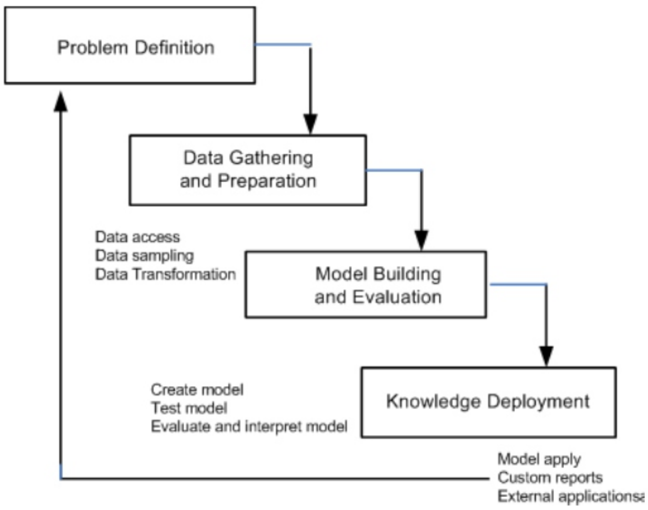

Fig. 1 Basic Data mining process

3. PREDICTING YIELD PRODUCTION

Yield prediction is a very important agricultural problem. Any farmer would like, in fact, to know, as soon as possible, how much yield he can expect. Attempts to solve this problem date back to the time when first farmers began to work soils in order to get profit. Since years, yield predictions have been performed by considering farmer's experience on particular fields and crops. However, this knowledge can also be obtained by exploiting information given by modern technologies, such as GPS. A multitude of sensor data can nowadays be relatively easily collected, so that farmers do not only harvest crops but also growing and growing amounts of data. These data are fine-scale, often highly correlated and carry spatial information. Figure 2 shows the final grouping.

Grouping	Clusters			Good ferm.	Bad ferm.	% problem
1	B			0	1	100
2	R	P		2	1	33
3	B	Br		1	0	0
4	R	Br		0	2	100
5	G	Br		0	2	100
6	P	Br		0	1	100
7	B	R		0	2	100
8	B	R	P	4	2	33
9	R	G	Br	1	1	50
10	R	P	G	1	0	0
11	R	P	Br	0	1	100
12	B	P	Br	0	1	100
13	B	R	Br	0	1	100

Fig. 2 Classification of wine fermentations using the k-means algorithm with k = 5 and by grouping the clusters in 13 groups

4. RECENT DEVELOPMENTS IN DATA MINING AND AGRICULTURE

The problem of predicting yield production can be solved by employing data mining techniques. Consider that sensor data are available for some time back to the past, where the corresponding yield productions have been recorded. All this information form a training set of data which can be exploited to learn how to classify future yield productions, once new sensor data are available. There are different data mining techniques that can be used for this purpose. In general, when considering the k-fold cross-validation technique, the original dataset can be divided in three parts: a training set, a validation set and a test set. Setting k equal to 10 or 20 is generally considered to be appropriate to remove bias. The regression model is trained on the training set until the prediction error on the validation set starts to rise. Once this happens, the training process is stopped and the error on the test set is reported for this fold. In spatial data, due to spatial autocorrelation, almost identical data records may end up in training, validation and test sets. The following table shows Data mining methodologies and its use in Agriculture domain.

Table: 1 Data mining methodologies and its use in Agriculture domain

Methodology	Applications
K-means	Forecasts of pollution in atmosphere Classifying soil in combination with GPS
k-nearest Neighbor	Simulating daily precipitations and other weather variable
Support Vector Machine	Analysis of different possible change of the weather scenario
Decision Tree Analysis	Prediction soil dept
Unsupervised Clustering	Generate cluster and determine any existence of pattern
WEKA Tool	Classification system for sorting and grading mushrooms

5. OTHER RECENT WORKS

We mention in this section some other recent interesting works in the field of data mining and agriculture. We begin with some other works related to the production of wine, which has been the focus of Section 2, where data mining approaches are employed for the prediction of problematic wine fermentations. The main aim of this work is to discover in advance fermentations that are going to be slow or stagnant, and to interfere with the process in order to guaranteeing a good fermentation. Other recent studies also concern the taste of the wine that is produced. In [4], for example, data mining techniques are employed in order to predict the taste of wine. This is done by creating a training set in which a classification of each sample (wine) is assigned by traditional wine tasters, that generally analyze some subjective parameters such as color, foam, flavor and savour of the wine. Once the classification task has been learned by exploiting the training set, data mining techniques are then supposed to substitute traditional wine tasters. Wine tastes are also analyzed in relation to seasonal climate effects.

6. CONCLUSIONS

This review presents a quick update with respect to the state-of-the-art in the field of data mining and agriculture. We mainly focus our attention on two particular problems. The first one is the problem of identifying problematic wine fermentations at the early stages of the process. A data mining approach to this problem has been discussed where the k-means algorithm was used. We described the recent developments on this problem, and in particular new studies where biclustering techniques are employed for identifying the compounds of wine that are most likely the cause of problematic fermentations. The second problem we consider is the one of predicting yield production. First approaches to this problem were based on standard data mining techniques, such as support vector regression and artificial neural networks. Recent works showed how to improve the quality of the classifications by employing the concept of spatial autocorrelation.

REFERENCES

1. S. Arivazhagan, R.N. Shebiah, S.S. Nidhyanandhan, L. Ganesan, Fruit Recognition using Color and Texture Features, Journal of Emerging Trends in Computing and Information Sciences 1(2), 90–94, 2010.
2. P. Baranowski, W. Mazurek, Detection of Physiological Disorders and Mechanical Defects in Apples using Thermography, International Agrophysics 23, 9–17, 2009.
3. S. Busygin, O.A. Prokopyev, P.M. Pardalos, Feature Selection for Consistent Biclustering via Fractional 0-1 Programming, Journal of Combinatorial Optimization 10, 7-21, 2005.
4. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Modeling Wine Preferences by Data Mining from Physicochemical Properties, Decision Support Systems 47(4), 547–553, 2009.
5. S. Cubero, N. Aleixos, E. Molt´o, J. G´omez-Sanchis, J. Blasco, Advances in Machine Vision Applications for Automatic Inspection and Quality Evaluation of Fruits and Vegetables, Food and Bioprocess Technology 4(4), 487–504, 2011.
6. L. Ding, J. Meng, Z. Yang, An Early Warning System of Pork Price in China Based on Decision Tree, IEEE Conference Proceedings, International Conference on E-Product E-Service and E-Entertainment (ICEEE), Henan, China, 1–6, 2010.
7. D.A. Griffith, Spatial Autocorrelation and Spatial Filtering, Advances in Spatial Science Series, Springer, New York, 2003.
8. M. Guarino, P. Jans, A. Costa, J-M. Aerts, D. Berckmans, Field Test of Algorithm for Automatic Cough Detection in Pig Houses, Computers and Electronics in Agriculture 62(1), 22–28, 2008.
9. D.S. Guru, Y.H. Sharath, S. Manjunath, Texture Features and KNN in Classification of Flower Images, International Journal of Computer Applications 1, Special Issue "Recent Trends in Image Processing and Pattern Recognition", 21–29, 2010.
10. R.P. Haff, Real-Time Correction of Distortion in X-ray Images of Cylindrical or Spherical Objects and its Application to Agricultural Commodities, Transactions of the American Society of Agricultural and Biological Engineers 51(1), 341–349, 2007.
11. R.P. Haff, N. Toyofuku, X-ray Detection of Defects and Contaminants in the Food Industry, Sensing and Instrumentation for Food Quality and Safety 2(4), 262–273, 2008.
12. J. Hartigan, Clustering Algorithms, John Wiles & Sons, New York, 1975. 13. M. Kovacevic, B. Bajat, B. Gajic, Soil Type Classification and Estimation of Soil Properties using Support Vector Machines, Geoderma 154(3–4), 340–347, 2010.
14. O.E. Kundakcioglu, P.M. Pardalos, The Complexity of Feature Selection for Consistent Biclustering, In: Clustering Challenges in Biological Networks, S. Butenko, P.M. Pardalos, W.A. Chaovalitwongse (Eds.), World Scientific Publishing, 2009.
15. A. Mucherino, A. Urtubia, Consistent Biclustering and Applications to Agriculture, IbaI Conference Proceedings, Proceedings of the Industrial Conference on Data Mining (ICDM10), Workshop "Data Mining in Agriculture" (DMA10), Berlin, Germany, 105-113, 2010.